Homework #13

1. In this problem you'll implement spectral clustering on the MNIST dataset.

   (a) Form a knn graph for the MNIST dataset. You may use sklearn to do this. Use your knn graph to from the graph Laplacian. $k = 30$ is a common choice.

   (b) Now using the first ℓ eigenvectors of the graph Laplacian that are not constant (remember here we want the smallest eigenvalues), embed the samples in $\mathbb{R}^\ell$ and apply kmeans with $k = 10$. You may use sklearn's kmeans function. Reuse your code from hw 11 to describe the image distribution in each cluster. What is the best ℓ? Do you do better than using k-means directly on the data?

2. In this problem, you'll provide the details for the Bayes optimal classifier results we discussed in class.

   (a) Define a probability distribution on (x, y) with $x \in \mathbb{R}^2$ and $y \in \{0, 1\}$ as follows. With probability $1/2$, $y = 0$ and $x \sim \mathcal{N}(0, I)$. (This means the pdf of x is given by,

   $$P(x) = \frac{\exp[-\|x\|^2/2]}{2\pi}) \tag{1}$$

   With probability $1/2$, $y = 1$ and $x \sim \mathcal{N}((\mu, 0), I)$. (This means the pdf of x is given by,

   $$P(x) = \frac{\exp[-((x_1 - \mu)^2 + x_2)^2/2]}{2\pi}) \tag{2}$$

   Determine the Bayes optimal classifier $\Phi(x) : \mathbb{R}^2 \to \{0, 1\}$ for this distribution.

   (b) Do **one** of the following.

   - Write a python function to sample (x_i, y_i) for $i = 1, 2, 3, \ldots, n$ from this distribution. Set $\mu = 1$. Numerically fit a logistic regression to the data and compare the decision boundary to the Bayes optimal decision boundary. Show that as you increase n, the logistic regression boundary approaches the Bayes optimal boundary. For some value of n, plot the data and the decision boundary. (You may use sklearn to fit the logistic regression. For sampling, you can use scipy's multivariate normal sampler.)

- Let $\ell(\alpha)$ be the log-likelihood of the logistic regression given n samples from the distribution. Show that the solution of $E[\nabla \ell(\alpha)] = 0$ gives a logistic regression decision boundary equal to the Bayes optimal boundary. (Given this result, a law of large numbers arguments shows convergence to the Bayes optimal boundary as $n \to \infty$.)

(c) Now consider a different distribution on (x, y). With probability 1/2, $y = 0$ and $x \sim \mathcal{N}(0, I)$, as in (a). With probability 1/2, $y = 1$ and $x \sim \mathcal{N}((10, 0), 5I)$ (This means the pdf of x is given by,

$$P(x) = \frac{\exp[-((x_1 - 10)^2 + x_2)^2/10]}{10\pi}) \tag{3}$$

Show that the decision boundary of the Bayes optimal classifier is not a line.

3. Recall homework 12, problem 3. In this problem we will solve the non-separable SVM problem introduced in that problem. The file `dataset_non_separable.csv` provides non separable data (x_i, y_i) for $i = 1, 2, \ldots, 2000$. $x_i \in \mathbb{R}^2$ and $y_i \in \{-1, 1\}$. For the kernel, use

$$k(x, y) = \exp[-\|x - y\|^2/.01] \tag{4}$$

Split the dataset into training and test datasets, each with 1000 samples. Recall the dual problem from hw 12, problem $3(a)$. For different values of C (the penalty parameter),

- Use qpsolver to compute the dual variable ν using the training dataset.
- Determine the accuracy of the fitted SVM for both the training and test dataset. Recall hw 12, problem $3(b)$ which shows how to predict y given ν.
- Plot the data with the points colored by cell type and then use Python's contour function to show the separating curve in $\mathbb{R}^2$.

Plot the trade-off between training and test accuracy as a function of C. What is the optimal C?